Meta-analysis of flow modeling performances - to build a matching system between catchment complexity and model types

Lu Zhuo[1*], Qiang Dai[1], Dawei Han[1*]

[1]WEMRC, Department of Civil Engineering, University of Bristol, Bristol, UK, BS8 1US

[*]Correspondence: lz7913@bristol.ac.uk and d.han@bristol.ac.uk

Abstract

Hydrological models play a significant role in modeling river flow for decision making support in water resource management. In the past decades, many researchers have made a great deal of efforts in calibrating and validating various models, with each study being focused on one or two models. As a result, there is a lack of comparative analysis on the performance of those models to guide hydrologists to choose appropriate models for the individual climate and physical conditions. This paper describes a two-level meta-analysis to develop a matching system between catchment complexity (based on catchment significant features CSFs) and model types. The intention is to use the available CSFs information for choosing the most suitable model type for a given catchment. In this study, the CSFs include the elements of climate, soil type, land cover and catchment scale. Specific choices of model types in small and medium catchments are further explored with all CSFs information obtained. In particular, it is interesting to find that semi-distributed models are the most suitable model type for catchments with the area over 3000km^2, regardless of other CSFs. The potential methodology for expanding the matching system between catchment complexity and model complexity is discussed.

Keywords: comparative assessment, river flow modeling, meta-analysis, hydrological model selection

1. Introduction:

Hydrological models play a significant role in simulation of river flow and decisions on water resource management. Similar to most science and engineering fields, the development of hydrological models has been unprecedented, especially during the past decades, which is largely driven by the advancement of computers, modern hydrological instruments, remote sensing technology, geographic information systems (GIS), digital elevation models (DEM), telecommunication networking facilities and so on. The growth may be viewed in term of benefits from the more detailed fields created within hydrology such as surface hydrology, subsurface hydrology, groundwater, forest hydrology, mountain hydrology etc. (Sivakumar et al., 2011), as well as the frequent reference of cross disciplinary theories and application of mathematical algorithms such as artificial neural networks, support vector machine and genetic algorithms within the hydrological community. Complex hydrological models such as physically based distributed models have a great deal of advantages over lumped models in describing spatially detailed hydrological processes. For example, they are capable of incorporating different kinds of spatially varied datasets such as soil type, land cover, geology, high resolution rainfall, temperature and other meteorological forcing inputs (Carpenter *et al.*, 2006). In addition, complex hydrological models are also capable of assessing pollutant and sediment movement (Anderson *et al.*, 1985). It is generally recognized that physically based distributed models (with the use of spatially varied catchment characteristics) may offer outputs with higher resolution and accuracy than lumped models (Beven, 1992, 2002; Boyle *et al.*, 2001; Smith *et al.*, 2004). Although hydrological models are moving towards a direction with more complex structure and mathematical sophistication, the selection of a suitable model/ model type for a given catchment becomes even harder than it used to be. As stated in Irstea (2014), "*It is difficult for a user to know which forecasting model to use ; each research institute develops its own model that they use for specifically defined purposes.*" Meanwhile, despite the advantages of complex hydrological models, there are two major concerns dominating discussions and debates on the current hydrologic studies: firstly, many hydrological models developed are often overly complicated, necessitating an enormous amount of parameters and require excessive data inputs which may not be available or needed and may even lead to over-fitting problems. As noted in Reed *et al.* (2004), the application of a fully distributed or a semi-distributed model may not improve flow modeling over lumped modeling. Secondly, models are often calibrated for a specific catchment (Irstea, 2014), and their extensions and generalizations to other catchment situations are rather difficult, especially the catchment classification is still in a state of infancy in current hydrology (Sivakumar *et al.*, 2011). Some studies have shown the usefulness of catchment classification with respect to hydrologic similarity, with examples such as McDonnell et al. (2004); Parajka et al. (2013); Salinas et al. (2013); Sawicz et al. (2011); Wagener et al. (2007). Winter (2001) introduced the idea of hydrologic landscapes, which were defined on the basis of similarity of climate, topography and geology, assuming that catchments that were similar with respect to these indicators would behave similarly in a hydrological sense. Bormann (2010) adopted a hydrologic classification system based on soil texture categories, assuming soil to be a major controlling factor of hydrologic similarity. In a similar manner, Ramachandra et al. (2006) explored catchments located within Indiana using physical features (area, channel length, channel slope etc.) to classify physically similar catchments. However, these conclusions are only constrained to the catchments studied and it is still puzzling to link the performance of a model to the physical characteristics of a catchment (Irstea, 2014). Furthermore no particular attention has been given to a systematic review of modeling performances in the published literature. Therefore the objective of this study is to review a considerable number of model developments based on the studies that have been published in peer-reviewed papers and compare their modeling performances in flow estimation. Many researchers have spent a great deal of efforts on calibrating and validating models, but each study only addressed one or two models. It is possible that there would be alternative models/ model types that may better suit a given catchment. Hence the aim of this paper is to review a large number of studies, with the initial step of learning from the differences and similarities between various catchments and between different hydrological models by using meta-analysis. After that suitable catchment features are selected to present catchment complexity and explore if any patterns would emerge between the categorized catchments and model types. Due to insufficient coverage of model types and catchment diversities in the existing publications, only a simple version of the matching system is presented in this paper. The rest of this paper is organized as follows: Section 2 describes the methodology and paper selection criteria applied in this study. Section 3 shows the data used and summary of the chosen studies for river flow modeling. Section 4 gives the result of the matching system between catchment features and model types. Discussion and conclusion are written in Section 5 and Section 6 respectively.

2. Methodology

In order to build the matching system between catchment complexity and model types, a two-level exploration based on the method used in Parajka *et al.* (2013) and Salinas *et al.* (2013) is applied. Since catchment significant features (CSFs) could be important indicators in discriminating catchment complexity as well as a way of choosing the right model type, so the initial step is to meta-analyze the model performance in terms of different model types and different CSFs by learning from their dissimilarities and similarities in a general way, and then to build a matching system accordingly (Section 2.1). The reasons for choosing climate region, soil type, land cover and catchment scale as CSFs are discussed in Section 2.2. The difference between model versions is discussed in Section 2.3.

2.1 Two Level Meta-analysis

Meta-analysis was initially proposed in the field of medicine (Antman *et al.*, 1992; DerSimonian *et al.*, 1986; Lau *et al.*, 1992), and then applied widely in engineering, science and management domains (Bork *et al.*, 2007; Duval *et al.*, 2000; Harsch *et al.*, 2009; King *et al.*, 2006). The benefit of using meta-analysis is that it covers a wide spectrum of possibilities (e.g. different models, various catchment characteristics) that go beyond what can be reasonably accomplished by a single case study. Generally speaking, there are five major steps towards a systematic literature survey in meta-analysis: framing the question, searching relevant publications, assessing study quality, summarizing the evidence and interpreting the findings (Khan *et al.*, 2003).

For the five steps introduced in the meta-analysis, the initial task is to choose suitable papers in the literature and then gather useful information to be used in this study. For paper selection, publications gathered from the international refereed journals are scrutinized for the results of river flow modeling. The databases of hydrological related disciplines (e.g. water resource management and flood risk management), ISI web of science and websites of hydrological organizations and environmental agencies with the relevant papers have been used, and various combinations of keywords such as ‘flow simulation’ and ‘hydrological model’ have been looked up. Furthermore, in order to improve the consistency of this study (e.g. the same performance indicator), several selection criteria are set for all the chosen papers to follow:

a) this study only chooses the papers with Nash –Sutcliffe efficiency (NSE) (Nash et al., 1970) as the indicator for model accuracy. This is because NSE is the most common and important performance measure used in hydrology. It is understood that modelers in other fields of environmental sciences are not often familiar with NSE, and NSE is not always effective in assessing a model’s performance (Schaefli et al., 2007). However this study is based on other researchers’ modeling results, with alternative performance criteria rarely reported. Therefore, only NSE could be considered in this study. As amply demonstrated by Diskin et al. (1977), there is no such index that is of universal application. In addition, as reported in Hall (2001), no one indicator is perfect in evaluating the complex hydrologic system. Nevertheless it is strongly suggested that future modelers should consider and provide more performance criteria for model evaluation, and the selection of indicators should depend on catchment characteristics and purposes of the models. The work carried out by Schaefli et al. (2005) is recommended, who adopted three performance criteria to assess the performance of a semi-lumped conceptual glacio-hydrological model, on three very complex high mountainous catchments.

b) papers with inappropriate model time interval are not included for assessment (the time interval should depend on the concentration time of a catchment, which is judged by the catchment scale), such as the studies carried out by Adamowski *et al.* (2010) and Saleh *et al.* (2004) who used daily time interval for small catchments (67km^2, 8-11km^2 respectively) which are not in line with other catchments.

c) this study only focuses on flow modeling with the whole hydrograph, hence some of the papers on real time flood forecasting are excluded, such as the modeling results given by Adamowski *et al.* (2010) and Lin *et al.* (2009).

For detailed information collection, since the idea of this study is to build a matching system between model complexity with various model types and catchment complexity with the identification based on CSFs, so comprehensive information on models, model types, temporal scales (e.g. flow modeling in hourly or daily time interval) and CSFs are required. In this study climate region, soil type, land cover and catchment scale are chosen as CSFs with the reasons discussed in the following section (Section 2.2). Moreover, it is observed that since there are a great variety of hydrological models used in the chosen studies, in order to generalize beyond individual studies, these models are grouped into four types as: black-box model, lumped model, semi-distributed model and fully-distributed model. Generally speaking lumped models are simpler with fewer parameters than fully distributed models. Therefore, the judgment is based on the number of parameters utilized in a model. It is known in Akaike Information Criterion (AIC) that a model is appraised by two components: one is how good a model is in fitting to the data and another is the number of parameters. Although it is true that model complexity can also depend on process complexity, there are few existing criteria to identify model process complexity in the literature, therefore adoption of the number of model parameters as model complexity is a pragmatic approach.

After gathering all the information about CSFs and model type of each flow modeling case, the catchment indicator (Ψ) for individual cases can then be represented as:

$$\Psi = \{CSF, M\} \tag{1}$$

$$CSF = \{SI, C, S, L\} \tag{2}$$

where M is model type with its subsets M1 (black-box), M2 (lumped), M3 (semi-distributed) and M4 (fully-distributed); SI is catchment scale with its subsets SI1 (small), SI2 (medium) and SI3 (large); C is climate type with its subsets C1 (continental), C2 (dry), C3 (mild temperate) and C4 (Tropical); S is soil type with its subsets S1 (clay), S2 (sand) and S3 (silt); L is land cover with its subsets L1 (urban), L2 (forest), L3 (agriculture) and L4 (grasslands). The detailed Ψ for each flow modeling case is presented in Table 4 in Section 4.3

After appraising the ability of each CSF on discriminating catchment complexity, two approaches of building the matching system between model types and catchment complexity are then carried out in Level 2 study. For the first attempt, the most suitable CSFs from Level 1 are classified into more detailed groups, which are then paired with model types. Alternatively, all CSFs are employed to represent catchment complexity and are then matched with suitable model types.

2.2 CSFs

The reasons of choosing climate region, soil type, land cover and catchment scale as CSFs are explained as:

- *Climate*: an important factor which dominates evaporation, rainfall and temperature of a catchment. It is noticed that there is inconsistency of climate definition between different papers. Some papers do not provide climate information at all, while others provide climate information but are based on unknown climate classification. For example, Chen et al. (2013)

describes the catchment as cold , dry climate which can either be arid or continental climate type; and Shi et al. (2011) defines the catchment as between northern subtropical and the warm temperate zone, which is rather difficult to summarize its main climate type. Therefore for the purpose of this study climate classification based on the Köppen climate classification system (Peel et al., 2007) is used, which is one of the most widely used climate classification systems. The classification combines information about temperature, precipitation, seasonality precipitation and native vegetation. Since precipitation represents water inflow, the temperature and vegetation represent evapotranspiration linked with energy and water, so they are representative of the water and energy balance. The Köppen classification includes five main groups as tropical, dry, mild temperate, continental and polar climates respectively with each having several subtypes. All the studies are categorized into the five main groups based on their locations and climate characteristics. However, two Swedish catchments in polar climate from the study of Lindström *et al.* (1997) have been removed, since they are the only two catchments from polar climate.

- *Soil type*: another factor that relates to flow generation. The variation of soil properties (e.g. partial size, porosity, and hydraulic conductivity) can affect the model performance. For the purpose of meta-analysis, the global soil texture map compiled by Webb *et al.* (2000) is used for soil classification. Because the dataset is developed to improve land surface hydrology parameterization, it has been widely applied for hydrological soil classification. The dataset indicates the top and bottom soil depths and the percentage of sand, silt and clay of each soil horizons in each of the 106 soil types over nine continents (NASA, 2000). Here, all the studies are categorized into the aforementioned three soil types according to the most abundant soil type at their catchment locations.
- *Land cover*: a significant catchment indicator. Different land covers such as urban, agriculture, forest can have different impacts on flow mechanism. Here, the land cover

classification based on the Global Land Cover map from SPOT VEGETATION data in 2000 GLC2000 (Bartholomé *et al.*, 2005) is adopted. The map has been used rather widely in environmental fields. The dataset divides land covers into seven major groups (forest, urban, wetlands, grasslands, agriculture, deserts, and snow & ice), with each having several subgroups. Among the studies that pass the aforementioned selection criteria, most are under urban, forest, agriculture and grasslands land covers, with only a few catchments spread over deserts and snow & ice land covers which are therefore removed from this study.

- *Catchment scale*: an indicator of the degree of aggregation of catchment processes related to scale effects, water storage within a catchment (Parajka et al., 2013), and catchment homogeneity because generally larger catchments tend to have more variations in land covers, soil types etc. And furthermore, an increase in the catchment area is often related to a decrease in the data availability and to the scale of the underlying information (Bormann *et al.*, 1999). However, there is no generally accepted definition for small, medium and large catchment sizes in the literature. For the purpose of Level 1 meta-analysis, the catchment boundary is defined by using Northern Ireland Water Framework Directive (2005) and Skoulikidis *et al.* (2006) as the references, with the small catchment size defined as 0 to 100km^2, the medium catchment size as 100 to 1000km^2 and the large catchment size as above 1000km^2.

It is important to mention that in this study only the dominant types of soil, land cover and climate are used. Although the data pattern/ smoothness/ heterogeneity are important, there are no easy ways of discriminating patterns and then link them with modeling outcomes. This is similar to principal component analysis (PCA) which is used to condense information into the first a few principal components; the dominant soil, land cover and climate enable the representation of the key relationships between them and modeling results.

2.3 Model versions

Since model versions are not reported in most literatures, especially for lumped and semi-distributed models, it is difficult to further assort the models according to different versions. Nevertheless it is still interesting to find out if there are any rising/ falling trends of the hydrological model performance with time in a general way, based on publication years of the studies (i.e. it is assumed that lately published papers have adopted higher version of models than papers published in earlier years). For this purpose, the plot of NSE results against the publication years of the selected literatures is shown in Figure 1a). It is surprising to see that there is a drop of the performance towards recent years, although the decrease is not remarkable. Moreover, the plots based on each individual model type are presented in Figure 1b) for a detailed investigation. It can be seen that the improvement of the black-box model is tremendous; however this result may not be representative due to its small sample and short time span of the cases included in this model type. For the other three model types, the fluctuations are negligible, with only an insignificant rise found in distributed model type and small decreases discovered in lumped and semi-distributed model types. Therefore, in conclusion the impact of model version is considered to be insignificant in this study, and it will not be included for the follow-on work.

3. Literatures and Data used

Table 1 lists the papers that pass the aforementioned selection criteria. It provides the summary information about study regions, models used, catchment characteristics and model efficiencies. The numbers of models and catchments in those papers vary. For example, some papers compare several hydrological models in one catchment, while other papers test one model in various catchments, which result in several model/ catchment assessments within one paper. As a result, a total of 119 assessments of flow modeling over 76 catchments form the foundation of Level 1 study.

It can be seen from Table 1 that most papers cover lumped and semi-distributed models (approximately 70% of the total papers) for flow modeling. Comparatively, a smaller number of the papers adopt fully-distributed models, which may be due to their complex model structure and vast data inputs. As in most fully-distributed models, flow generation is only a part of the system functions. As a result, they are used more widely in simulating water quality and sediment transport (Ahearn *et al.*, 2005; Konz *et al.*, 2011; Refshaard *et al.*, 1995), and with the coupling between water balance and energy balance for climate change investigations (Christensen *et al.*, 2004; Lin *et al.*, 2009).

4. Results

In order to build a matching system between catchment complexity and model types, assessments on each CSF and each model type must be carried out firstly which are presented in Section 4.1 (Level 1). And catchment are then classified under two schemes: firstly by choosing the most significant CSF as shown in Section 4.2 (Level 2), and secondly by adopting all CSFs information as presented in Section 4.3 (Level 2).

4.1 Comparative assessment of model types and CSFs (Level 1)

4.1.1 Model types

The model types used in Level 1 study include black-box, lumped, semi-distributed and fully-distributed models. Figure 2 illustrates the NSE performances in terms of different model types. It can be seen that the performance of lumped models are the best, followed by semi-distributed models. Although the general accuracy of semi-distributed models is good, there are still 5 studies that performed below 0.40 NSE. The excellent performance of lumped models may be due to their wide application and long history; hence they are relatively mature technique in hydrologic modeling and are also easier to be calibrated. In contrast, it is surprising to see that the fully-distributed models perform the second last with the mean NSE only at 0.58, which could be explained by the following three reasons. Firstly, fully-distributed models are comprehensive systems which couple both water and energy balances, hence they require a large amount of data and catchment information. Therefore, if the input data is not sufficient, then the model efficiency will clearly not be as good as expected. Secondly, since fully-distributed models have complex structures with many parameters, they tend to suffer from over-fitting problems. Thirdly, since fully-distributed models are more complex than lumped models, they require modelers to possess high level modeling skills and experiences; hence such a model type depends more on subjective factors. Among all models, black-box models perform the worst with the mean NSE only at 0.49. Because models in this type do not consider the physical meaning of hydrological processes, they could be easily misunderstood and poorly calibrated.

4.1.2 Climate type

Climate, as an important factor affecting evapotranspiration and precipitation, can be expected to influence the performance of flow modeling. The summary of climate types and corresponding main climate groups are presented in Table 2. As illustrated in Figure 3, the synthesis of the chosen studies indicates that most of the studies are carried out in Europe and USA, and majority studies are carried out in mild temperate climate rather than in dry and tropical climates. The results of NSE performances with respect to the four climate types are presented in Figure 4a), which shows that the average performance of flow modeling tends to be lower in mild temperate climate than in dry and tropical climates. For mild temperate climate, the range of NSE changes from below 0.30 (Bell et al., 2001; Saleh et al., 2004; Smith et al., 2004; Xevi et al., 1997) to higher than 0.90 (Bell et al., 2001; Lindström et al., 1997), with 14 out of 92 studies exceeding 0.80. This is because more catchments are studied in locations covered with mild temperate climate; hence more diverse catchment situations are included, which lead to higher variations of the performance results. It is surprising to see that the performances of flow modeling in dry areas are quite stable, especially with the high accuracy (NSE=0.81) observed in Senthil et al. (2005). Similar results are obtained in tropical climate, with the mean NSE at 0.84, especially in the study carried out by Campling et al. (2002), who used a semi-distributed TOPMODEL in River Ebonyi, Nigeria and obtained almost 0.90 NSE. However, the results for both dry and tropical climates may not be representative due to the small number of cases included in both types.

4.1.3 Soil type

Soil plays a significant role in the hydrological cycle, because various soil properties can affect the formation of runoff. The assessment of NSE performances with respect to the three soil types is presented in Figure 4b). It is clear that the performances of flow modeling carried out in silt based catchments are the best with the mean NSE value of 0.76 and also with the narrowest NSE range between 0.49 and 0.91. Such catchments appear mostly in mid-USA (such as Iowa (Tokar *et al.*, 2000), Illinois, Indiana (Singh *et al.*, 2005), Oklahoma (Khan, 1993; Yew *et al.*, 1997)), Europe (such as Italy (Todini, 1996) and England (Bell *et al.*, 2001)). Comparatively, in sandy catchments the efficiencies of the applied models are reasonably sparse, with an appreciable difference between the best performed case (0.91 NSE, by the semi-distributed HBV-96 model in a Swedish sandy catchment by Lindström *et al.* (1997) and the worst performed case (0.24 NSE, by the fully-distributed MIKE-SHE model in Neuenkirchen research catchment, Germany by Xevi *et al.* (1997). In conclusion, the performances of flow modeling in silt catchments are more stable than in sandy and clayey catchments. This is because the catchments covered with silt soil type are often located in humid areas, and hydrological models normally perform more stable in these locations than in arid and cold catchments.

4.1.4 Land cover

In addition to the impact of soil on flow modeling, land cover is also influential. As shown in Figure 4c), the forest group performs the worst among all land covers. It also has the largest NSE difference between the worst and the best cases. Comparatively, the performance of flow modeling in grasslands is more efficient. As mentioned in Section 2, the influence of land cover is a composite indicator including a range of processes. For example, catchments covered with intensive vegetation like forests are sophisticated systems, so researchers tend to use complex distributed models. However, lacking of the physical measurements of individual parameters or arbitrarily fixing certain parameters can result in significant errors. For example, the results presented in 'the distributed model intercomparison project' (Reed *et al.*, 2004; Smith *et al.*, 2004) are extremely low (NSE range between -0.26 and 0.27), and it is interesting to find that most of these poor results are from the SWAT model. Compared with the extreme situation of the forest group, the performances of flow modeling in urban and agriculture are more acceptable with the mean NSE of 0.70 and 0.70 respectively.

4.1.5 Catchment scale

Catchment scale may be a useful indicator of catchment homogeneity. The results in Figure 4d) present a rather clear increase trend of the efficiency with catchment scale for all the studies. The mean NSE performance is 0.39 in small catchments and rises to approximately 0.70 for large catchments. In addition, the precision rises when catchment area increases. For example, the standard deviation deceases dramatically from 0.76 for small catchments to only 0.18 for large catchments. This pattern of the performance with catchment scale may be due to the following reason: when the catchment area increases, some of the hydrological variability is averaged out due to an interplay of hydrological processes both spatially and temporally, hence the flow modeling is improved. Similar results are revealed in Merz *et al.* (2009, 2011) and Nester *et al.* (2011), and an ungauged catchment study (Parajka *et al.*, 2013). Compared with climate, soil and land cover, catchment scale shows stronger evidence as an indicator of catchment complexity. Therefore, catchment scale would be used to further explore the detailed catchment complexity in Section 4.2.

4.2 Catchment scale match with model types (Level 2)

Considering the reason explained in Section 4.1.5, further exploration is implemented to examine the ability of catchment scale in representing the catchment complexity, and then to build a matching system between catchment complexity and model types. In this study, the catchment area varies from 0.36km^2 to 795,500km^2; with such big differences, one model type is clearly incapable of covering all catchment scales. Hence in order to find the most suitable model types for various catchment scales, a correlation between catchment areas and model performances is explored in respect to the four model types. As shown in Figure 5, there is an evident elevation of performances across all model types when the catchment expands. Furthermore, it is found that the general performance of fully-distributed models is unsatisfactory with their majority NSE efficiencies lower than 0.80. However, it is noted that when the results of all model types are plotted in one figure, it is difficult to determine which model type is better for which catchment scale. The reason is because, semi-distributed models are widely used in large catchments, however this does not yield to the conclusion that they are better in large catchments, as the number of studies for lumped models in large catchments is too small to compare. In addition, a similar situation is found for lumped models in medium catchments. Therefore, in order to avoid the impacts of this preferred bias for catchment scale, cases are further divided into small, medium and large catchments. In Level 1 meta-analysis, only a rough definition of catchment size is adopted. In order to link between different catchment sizes and suitable model types, the border lines between small and medium catchments, and between medium and large catchments need to be discovered. Ultimately, the optimal boundary should reflect a clear pattern between the catchment sizes and the corresponding suitable model types, and the change of the border lines should have an appreciable impact on the pattern. For this purpose, the trial and error method is applied to discover a suitable definition of catchment size groups. The border line between small and medium catchments is tuned firstly while the boundary between medium and large catchments is kept unchanged (use 50km^2 as changing steps). Until the pattern in small catchments is clear, the border line between medium and large catchments would then be tuned by applying the same method (with gradually increased changing steps from 50km^2 to 1000km^2). The final boundary is discovered as: small catchment between 0 and 200km^2; medium catchment between 200 and 3000km^2 and large catchment greater than 3000km^2. This result agrees with the definition in Mnatsakanyan *et al.* (2007).

Figure 6 presents the performances of flow modeling in terms of different catchment size groups. Since the number of studies using black-box models are too small, these results have been excluded in this part. Meanwhile, it is worth reaffirming that the results in Figure 6 are not shown in the order of catchment scales; instead they are sorted in the order of ascending NSEs. In the results presented in Figure 6, it is convincing to say that semi-distributed models give better modeling performance in large catchments. On the other hand, in small and medium catchments, there is no distinctive difference between model types, with only slightly better performance observed for lumped models in small catchments and semi-distributed models in medium catchments. In addition, it is obvious that there is still large disparity within each model type. For example, with medium catchments, some of the lumped models surpass the semi-distributed models with efficiency as high as 0.85, i.e. Tank and Sacramento models used in Smith *et al* (2004) and Todini (1996) and similar results are discovered in other cases (Boyle *et al.*, 2001; Carpenter *et al.*, 2001; Koren *et al.*, 2004; Smith *et al.*, 1999; Vieux *et al.*, 2003; Zhang *et al.*, 2004). Therefore, it is necessary to perform a more specific classification within each model type. For this purpose, model names are used to stratify models in preferred and non-preferred (with 0.80 NSE as the threshold, because modelers lose interest when the result is below 0.8) groups as shown in Table 3. As discussed in Section 2.3, model version is not considered in this result. It can be seen from Table 3 that SAC-SMA is not as suitable as Midlands Catchment Runoff Model (MCRM) in small catchments. For medium and large catchments, the SWAT model is not efficient in both catchment sizes; comparatively, HBV model is better. The summary of Table 3 can be used as a simple matching system when other CSFs are deficient, especially for catchments with areas over 3000km^2.

4.3 All CSFs match with model types (Level 2)

It can be seen from Section 4.2 that the disparity among three model types in small and medium catchments are still not as distinctive as the one found in large catchments. However, if all the CSFs information of a catchment is available, the matching system between catchment complexity and model types can be further developed based on the extra information. Table 4 presents the Ψ of all the studies, which corresponds to Table 1. Similar to the threshold set in the previous section, in order to enhance the reliability of the matching system, only studies with NSE performance greater than 0.80 are taken into consideration. Moreover, it is noted that the catchment size in this section is categorized based on the boundary found in Section 4.2. For the methodology, for catchments with identical CSF codes, their NSE performances via different model types are compared, and the model type with better accuracy is specified as the preferred model type. The matching system between catchment complexity and model complexity from this scheme is presented in Table 5. It can be seen that the efficiently performed semi-distributed models occupy most of the preferred model group, as well as the widest catchment cases (with 10 out of 19 CSFs situations). The performance is followed by lumped models (with 6 out of 19 CSFs situations). This matching system emphasizes the conclusion made in Section 4.2 that for large catchments (SI3), semi-distributed models (M3) are the optimal model choice regardless of other CSFs conditions.

5. Discussion

Catchment complexity and model complexity (model types in this study) are tightly associated with each other, and with the catchment complexity as the most important one. Therefore, until the catchment complexity is well defined, it will be difficult to provide model selection guidance.

Firstly, how to measure the catchment complexity? As presented in Section 4.2 the size of a catchment is a way of measuring catchment complexity. However it should not be the sole indicator and there are other possible alternatives. As shown in Figure 6, using catchment size as a sole indicator in small and medium catchments is clearly insufficient. In order to improve the matching system, all CSFs are employed in the second approach as shown in Section 4.3. Although the results in Table 5 remedy the deficiency occurred in small and medium catchments, they do not cover the whole possibilities. There are some studies on catchment complexity with attempts to find guidelines to classify catchments into different groups, such as the studies carried out by Reed (1999); Sivakumar *et al.* (2011); Troch *et al.* (2013). However, these studies do not address directly the catchment complexity in relation to CSFs and are only preliminary and limited. Therefore measuring the catchment complexity is still an unsolved problem in hydrology.

Secondly, how to deal with the model complexity? In the current literature, there is no generally accepted definition for the complexity of hydrological models and one way of measuring model complexity is by counting model parameters such as the AIC formula. Generally speaking, hydrological models are vaguely catalogued into three types: lumped, semi-distributed and fully-distributed models (with lumped models having the least number of parameters as compared with the fully distributed models which divide a catchment into thousand or even hundred thousand grids, with each grid having its own set of parameters); Somehow it is true that model complexity can also depend on process complexity and the parameter counting method is not always working in hydrologic modeling. Because a model with high process complexity may be lumped, a semi-distributed model may have a lot of insensitive parameters, and a fully-distributed model with a vast number of required parameters may have many parameters fixed for all the grids. However as there are few existing agreed criteria to evaluate process complexity, a better agreed system is urgently desired, probably with a weighting methodology (e.g. sensitive parameters should have higher weightings while less sensitive parameters should have less weightings). Sensitivity analysis could be useful in assigning weightings in this case.

Apart from the aforementioned challenges in measuring catchment complexity and model complexity, the data sufficiency issue is also a big challenge that need to be solved in hydrology (e.g. how much data information is sufficient for a catchment). Somehow in order to solve these problems, a physically based fully distributed hydrological model as virtual catchments might be employed to tackle this challenge. Since they are virtual catchments, it will be rather easy to generate thousand or even hundred thousand by the Monte-Carlo method. A large number of cases could be generated to link catchment complexity with model complexity. In addition, the skill of a modeler also plays an important role. However it is impossible to judge modelers' skill from their results published in international refereed journal papers. Since this study is based on peer-reviewed results, it is assumed that all modelers have adequate modeling skills. Albeit they do not, there are no appropriate ways to judge. Nevertheless there may be a way to tackle this issue , for example by a controlled study carried out between different groups of modelers with different skill levels, working on the same catchments and same datasets. However such studies have rarely been carried out and reported in the literature. It will be an interesting study to be implemented by the hydrological research community in the future.

Moreover, could a multi-hydrological-model approach help improving flow simulations and forecasts, and thereby obtain more accurate outputs? Logically, the answer could be yes. A study carried out by the PREMHYCE project (PREMHYCE) led by Irstea has proven that a combination of five French operational hydrological models works more effectively than the models used individually. Recent advancement in Data Mining technology such as Fuzzy logic, Logistic Regression and Random Forest may also be used as a combination of models to simulate the complicated river systems.

6. Conclusion

Hydrologic modeling has become an important research domain in hydrology, particularly facilitated by the fast development of computing technology and mathematical algorithms. Hydrological models are increasingly complex and computationally intensive. Therefore, to choose a suitable model/model type for a given catchment is becoming a complex problem. Because there are some fundamental issues remain unsolved in the hydrological community such as what is the optimum way of measuring catchment complexity and how to discriminate various hydrological models (with parameter counting, or alternative ways). Hence the motivation behind this study is to gather a large number of publications with different types of hydrological models and various catchment situations, to build a preliminary matching system between catchment complexity and model types. We hope it will provide useful guidance for future hydrologists to assess various catchments and choose different hydrological models.

In this study, a significant correlation between semi-distributed models and large catchments (>3000km^2) are uncovered. Moreover, in order to improve the results in small and medium catchments, the relationships in terms of all CSFs are taken into further exploration. These two attempts provide specific choices of models and model types across a wide range of catchment scales. Since the hydrological community has already spent a great deal of effort and time in utilizing various hydrological models with a variety of catchments, a clear matching system between catchment complexity and model complexity is urgently needed. Several recent studies (Sawicz *et al.*, 2011; Sivakumar *et al.*, 2011; Troch *et al.*, 2013; Wagener *et al.*, 2007) have attempted to classify catchment complexity based on hydrologic similarity; however an overarching method based on multiple factors such as CSFs. has yet to emerge. We hope that this study will be a useful step towards further engaging the hydrological community in advancing the research of matching catchment complexity and model complexity for better flow modeling and forecasting.

Acknowledgements

The first author is grateful for the financial support from Bristol University Alumni Foundation in her PhD project. The digital maps of the catchment information in this study is kindly provided by University of Veterinary Medicine Vienna (VUW) for the Köppen climate classification; NASA Earth Data for the global soil texture, and Land Resource Management Unit of the European

Commission for the Global Land Cover 2000.

References:

Ackerman, D., Schiff, K. C., & Weisberg, S. B. (2005). EVALUATING HSPF IN AN ARID, URBANIZED WATERSHED1. *JAWRA Journal of the American Water Resources Association, 41*(2), 477-486.

Adamowski, J., & Sun, K. (2010). Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. *Journal of Hydrology, 390*(1), 85-91.

Ahearn, D. S., Sheibley, R. W., Dahlgren, R. A., Anderson, M., Johnson, J., & Tate, K. W. (2005). Land use and land cover influence on water quality in the last free-flowing river draining the western Sierra Nevada, California. *Journal of Hydrology, 313*(3), 234-247.

Ambroise, B., Freer, J., & Beven, K. (1996). Application of a generalized TOPMODEL to the small Ringelbach catchment, Vosges, France. *Water resources research, 32*(7), 2147-2159.

Anderson, M. G., & Burt, T. P. (1985). *Hydrological forecasting* (Vol. 372): Wiley Chichester.

Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA, 268*(2), 240-248.

Bartholomé, E., & Belward, A. (2005). GLC2000: a new approach to global land cover mapping from Earth observation data. *International Journal of Remote Sensing, 26*(9), 1959-1977.

Bell, V., Carrington, D., & Moore, R. (2001). Comparison of rainfall-runoff models for flood forecasting. Part 2: Calibration and evaluation of models.

Beven, K. (1992). Editorial: Future of distributed modelling. *Hydrological processes, 6*(3), 253-254.

Beven, K. (2002). Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system. *Hydrological processes, 16*(2), 189-206.

Bork, E. W., & Su, J. G. (2007). Integrating LIDAR data and multispectral imagery for enhanced classification of rangeland vegetation: A meta analysis. *Remote Sensing of Environment, 111*(1), 11-24.

Bormann, H. (2010). Towards a hydrologically motivated soil texture classification. *Geoderma, 157*(3), 142-153.

Bormann, H., Diekkrueger, B., & Richter, O. (1999). Effects of spatial data resolution on the calculation of regional water balances. *IAHS Publication(International Association of Hydrological Sciences)*(254), 193-202.

Boyle, D. P., Gupta, H. V., Sorooshian, S., Koren, V., Zhang, Z., & Smith, M. (2001). Toward improved streamflow forecasts: Value of semidistributed modeling. *Water resources research, 37*(11), 2749-2759.

Cabus, P. (2008). River flow prediction through rainfall–runoff modelling with a probability-distributed model (PDM) in Flanders, Belgium. *Agricultural Water Management, 95*(7), 859-868.

Campling, P., Gobin, A., Beven, K., & Feyen, J. (2002). Rainfall-runoff modelling of a humid tropical catchment: the TOPMODEL approach. *Hydrological processes, 16*(2), 231-253.

Carpenter, T. M., & Georgakakos, K. P. (2006). Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. *Journal of Hydrology, 329*(1), 174-185.

Carpenter, T. M., Georgakakos, K. P., & Sperfslagea, J. A. (2001). On the parametric and NEXRAD-radar sensitivities of a distributed hydrologic model suitable for operational use. *Journal of Hydrology, 253*(1), 169-193.

Chen, X., Yang, T., Wang, X., Xu, C.-Y., & Yu, Z. (2013). Uncertainty Intercomparison of Different Hydrological Models in Simulating Extreme Flows. *Water Resources Management*, 1-17.

Christensen, N. S., Wood, A. W., Voisin, N., Lettenmaier, D. P., & Palmer, R. N. (2004). The effects of climate change on the hydrology and water resources of the Colorado River basin. *Climatic Change, 62*(1-3), 337-363.

Cuo, L., Lettenmaier, D. P., Mattheussen, B. V., Storck, P., & Wiley, M. (2008). Hydrologic prediction for urban watersheds with the Distributed Hydrology–Soil–Vegetation Model. *Hydrological processes, 22*(21), 4205-4213.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials, 7*(3), 177-188.

Diskin, M., & Simon, E. (1977). A procedure for the selection of objective functions for hydrologic simulation models. *Journal of Hydrology, 34*(1), 129-149.

Duval, S., & Tweedie, R. (2000). Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics, 56*(2), 455-463.

Hall, M. (2001). How well does your model fit the data? *Journal of Hydroinformatics, 3*, 49-55.

Harsch, M. A., Hulme, P. E., McGlone, M. S., & Duncan, R. P. (2009). Are treelines advancing? A global meta-analysis of treeline response to climate warming. *Ecology letters, 12*(10), 1040-1049.

He, S., Nan, Z., Wang, S., & Wang, Y. (2012). 四个概念性水文模型在黑河流域上游的应用与比较分析. *JOURNAL OF CHINA HYDROLOGY, 32*(3).

Im, S., Kim, H., Kim, C., & Jang, C. (2009). Assessing the impacts of land use changes on watershed hydrology using MIKE SHE. *Environmental geology, 57*(1), 231-239.

irstea. (2014). Hydrological forecasting models: towards a better anticipation of low water flows? Retrieved 29/07/2014, from Irstea National Research Institute of Science and Technology for Environment and Agriculture http://www.irstea.fr/en/all-news/water-department/hydrological-forecasting-models-anticipation-low-water

Khan. (1993). Xinanjiang Model on Bird Creek catchment in USA. *Pakistan Journal of Agricultural Research, 14*(4), 373-382.

Khan, K., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine, 96*(3), 118-121.

Khan, M. (1993). Xinanjiang Model on Bird Creek catchment in USA. *Pakistan Journal of Agricultural Research, 14*(4), 373-382.

King, W. R., & He, J. (2006). A meta-analysis of the technology acceptance model. *Information & Management, 43*(6), 740-755.

Konz, M., Chiari, M., Rimkus, S., Turowski, J., Molnar, P., Rickenmann, D., & Burlando, P. (2011). Sediment transport modelling in a distributed physically based hydrological catchment model. *Hydrology and Earth System Sciences, 15*(9), 2821-2837.

Koren, V., Reed, S., Smith, M., Zhang, Z., & Seo, D.-J. (2004). Hydrology laboratory research modeling system (HL-RMS) of the US national weather service. *Journal of Hydrology, 291*(3), 297-318.

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med, 327*(4), 248-254.

Lin, G.-F., Chen, G.-R., Huang, P.-Y., & Chou, Y.-C. (2009). Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods. *Journal of Hydrology, 372*(1), 17-29.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology, 201*(1), 272-288.

McDonnell, J. J., & Woods, R. (2004). On the need for catchment classification. *Journal of Hydrology,*

299(1), 2-3.
Merz, R., Parajka, J., & Blöschl, G. (2009). Scale effects in conceptual hydrological modeling. *Water resources research, 45*(9), W09405.
Merz, R., Parajka, J., & Blöschl, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. *Water resources research, 47*(2), W02531.
Mnatsakanyan, B., Aghababyan, K., & Chilingaryan, L. (2007). On the Calculation of Maximal Outlets of Small Mountainous Rivers (in Armenian conditions). *NATO SECURITY THROUGH SCIENCE SERIES D-INFORMATION AND COMMUNICATION SECURITY, 12*, 245.
NASA. (2000). GLOBAL SOIL TEXTURE AND DERIVED WATER-HOLDING CAPACITIES (WEBB ET AL.). Retrieved 29/07/2014 http://daac.ornl.gov/SOILS/guides/Webb.html
Nash, J., & Sutcliffe, J. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology, 10*(3), 282-290.
Nester, T., Kirnbauer, R., Gutknecht, D., & Blöschl, G. (2011). Climate and catchment controls on the performance of regional flood simulations. *Journal of Hydrology, 402*(3), 340-356.
Northern Ireland Water Framework Directive. (2005). River Reference Conditions, Article 5 Characterization Technical Report: Northern Ireland Water Framework Directive.
Parajka, J., Viglione, A., Rogger, M., Salinas, J., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins–Part 1: Runoff hydrograph studies. *Hydrology and Earth System Sciences Discussions, 10*(1), 375-409.
Pechlivanidis, I., McIntyre, N., & Wheater, H. (2010). Calibration of the semi-distributed PDM rainfall–runoff model in the Upper Lee catchment, UK. *Journal of Hydrology, 386*(1), 198-209.
Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences Discussions, 4*(2), 439-473.
PREMHYCE. (2011-2013). Prediction of Low-Flow Periods by Hydrological Models: Comparison and Evaluation, financed by ONEMA and MEDDE. Scientific partners: Irstea, BRGM, Météo France, EDF-DTG, University of Lorraine.
Ramachandra Rao, A., & Srinivas, V. (2006). Regionalization of watersheds by hybrid-cluster analysis. *Journal of Hydrology, 318*(1), 37-56.
Reed, Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., & DMIP, P. (2004). Overall distributed model intercomparison project results. *Journal of Hydrology, 298*(1), 27-60.
Reed, D. (1999). Flood estimation handbook. 1 Overview. *Institute of Hydrology, UK*.
Refshaard, J., Storm, B., & Singh, V. (1995). MIKE SHE. *Computer models of watershed hydrology.*, 809-846.
Rouhani, H., Willems, P., Wyseure, G., & Feyen, J. (2007). Parameter estimation in semi-distributed hydrological catchment modelling using a multi-criteria objective function. *Hydrological processes, 21*(22), 2998-3008.
Saleh, A., & Du, B. (2004). Evaluation of SWAT and HSPF within BASINS program for the upper North Bosque River watershed in central Texas. *Transactions of the ASAE, 47*(4), 1039-1049.
Salinas, J., Laaha, G., Rogger, M., Parajka, J., Viglione, A., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins--Part 2: Flood and low flow studies. *Hydrology & Earth System Sciences Discussions, 10*(1).
Sawicz, K., Wagener, T., Sivapalan, M., Troch, P., & Carrillo, G. (2011a). Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences, 15*(9), 2895-2911.
Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes, 21*(15), 2075-2080.
Schaefli, B., Hingray, B., Niggli, M., & Musy, A. (2005). A conceptual glacio-hydrological model for high mountainous catchments. *Hydrology and earth system sciences,*

9(HYDRAM-ARTICLE-2005-006), 95-109.
Senthil, K., Sudheer, K., Jain, S., & Agarwal, P. (2005). Rainfall-runoff modelling using artificial neural networks: comparison of network types. *Hydrological processes, 19*(6), 1277-1291.
Shi, P., Chen, C., Srinivasan, R., Zhang, X., Cai, T., Fang, X., . . . Li, Q. (2011). Evaluating the SWAT model for hydrological modeling in the Xixian watershed and a comparison with the XAJ model. *Water Resources Management, 25*(10), 2595-2612.
Singh, J., Knapp, H. V., Arnold, J., & Demissie, M. (2005). Hydrological modeling of the iroquois river watershed using HSPF and SWAT1. *JAWRA Journal of the American Water Resources Association, 41*(2), 343-360.
Sivakumar, B., & Singh, V. (2011). Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework. *Hydrology and Earth System Sciences Discussions, 8*(3), 4427-4458.
Skoulikidis, N. T., Amaxidis, Y., Bertahas, I., Laschou, S., & Gritzalis, K. (2006). Analysis of factors driving stream water composition and synthesis of management tools—a case study on small/medium Greek catchments. *Science of the total environment, 362*(1), 205-241.
Smith, M. B., Koren, V., Finnerty, B., & Johnson, D. L. (1999). *Distributed modeling: Phase 1 results*: US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service.
Smith, M. B., Seo, D.-J., Koren, V., Reed, S., Zhang, Z., Duan, Q., . . . Cong, S. (2004). The distributed model intercomparison project (DMIP): motivation and experiment design. *Journal of Hydrology, 298*(1), 4-26.
Takeuchi, K., Hapuarachchi, P., Zhou, M., Ishidaira, H., & Magome, J. (2008). A BTOP model to extend TOPMODEL for distributed hydrological simulation of large basins. *Hydrological processes, 22*(17), 3236-3251.
Todini, E. (1996). The ARNO rainfall—runoff model. *Journal of Hydrology, 175*(1), 339-382.
Tokar, A. S., & Markus, M. (2000). Precipitation-runoff modeling using artificial neural networks and conceptual models. *Journal of Hydrologic Engineering, 5*(2), 156-161.
Troch, P. A., Carrillo, G., Sivapalan, M., Wagener, T., & Sawicz, K. (2013). Climate-vegetation-soil interactions and long-term hydrologic partitioning: signatures of catchment co-evolution. *Hydrology and Earth System Sciences Discussions, 10*(3), 2927-2954.
Vieux, B. E., & Moreda, F. G. (2003). Ordered physics-based parameter adjustment of a distributed model. *Water Science and Application, 6*, 267-281.
Vischel, T., Pegram, G., Sinclair, S., & Parak, M. (2008). Implementation of the TOPKAPI model in South Africa: Initial results from the Liebenbergsvlei catchment. *Water Sa, 34*(3), 1-12.
Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment classification and hydrologic similarity. *Geography Compass, 1*(4), 901-931.
Webb, R. W., Rosenzweig, C. E., & Levine, E. R. (2000). Global Soil Texture and Derived Water-Holding Capacities (Webb et al.). *Data set. Available on-line [http://www.daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA*.
Whitaker, A., Alila, Y., Beckers, J., & Toews, D. (2003). Application of the distributed hydrology soil vegetation model to Redfish Creek, British Columbia: Model evaluation using internal catchment data. *Hydrological processes, 17*(2), 199-224.
Winter, T. C. (2001). THE CONCEPT OF HYDROLOGIC LANDSCAPES1: Wiley Online Library.
Xevi, E., Christiaens, K., Espino, A., Sewnandan, W., Mallants, D., Sørensen, H., & Feyen, J. (1997). Calibration, validation and sensitivity analysis of the MIKE-SHE model using the Neuenkirchen catchment as case study. *Water Resources Management, 11*(3), 219-242.
Xie, H., & Lian, Y. (2013). Uncertainty-based evaluation and comparison of SWAT and HSPF applications to the Illinois River Basin. *Journal of Hydrology, 481*, 119-131.
Xiong, L., & Guo, S. (2004). Effects of the catchment runoff coefficient on the performance of

TOPMODEL in rainfall–runoff modelling. *Hydrological processes, 18*(10), 1823-1836.

Yew Gan, T., Dlamini, E. M., & Biftu, G. F. (1997). Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. *Journal of Hydrology, 192*(1), 81-103.

Zhang, Koren, V., Smith, M., Reed, S., & Wang, D. (2004). Use of next generation weather radar data and basin disaggregation to improve continuous hydrograph simulations. *Journal of Hydrologic Engineering, 9*(2), 103-115.

Zhang, Z., Wang, S., Sun, G., McNulty, S. G., Zhang, H., Li, J., . . . Strauss, P. (2008). Evaluation of the MIKE SHE Model for Application in the Loess Plateau, China1. *JAWRA Journal of the American Water Resources Association, 44*(5), 1108-1120.

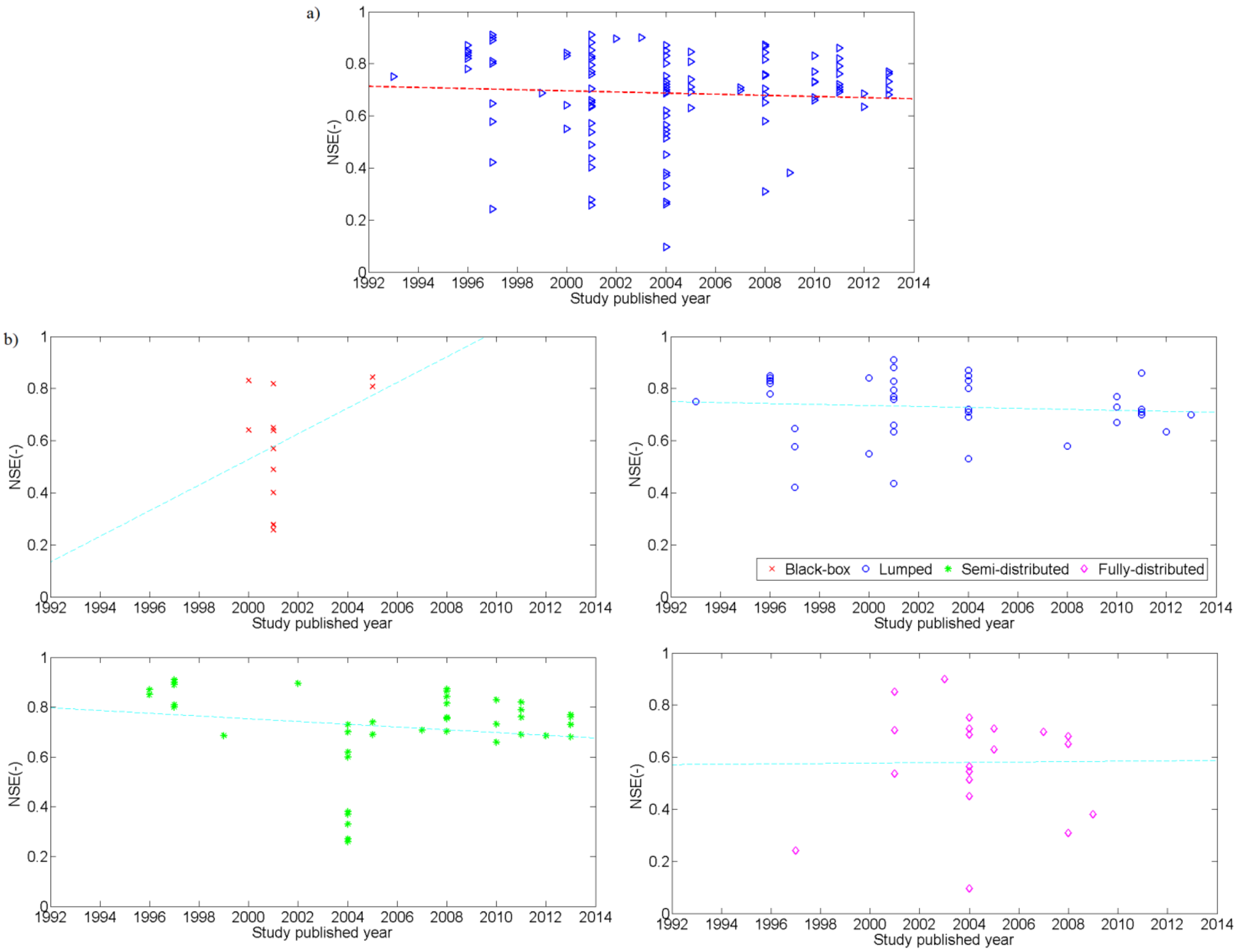

Figure 1. NSE performances against publication years of all the selected assessments. a) for all model types, b) for four individual model types with black-box, lumped, semi-distributed and fully distributed model types respectively.

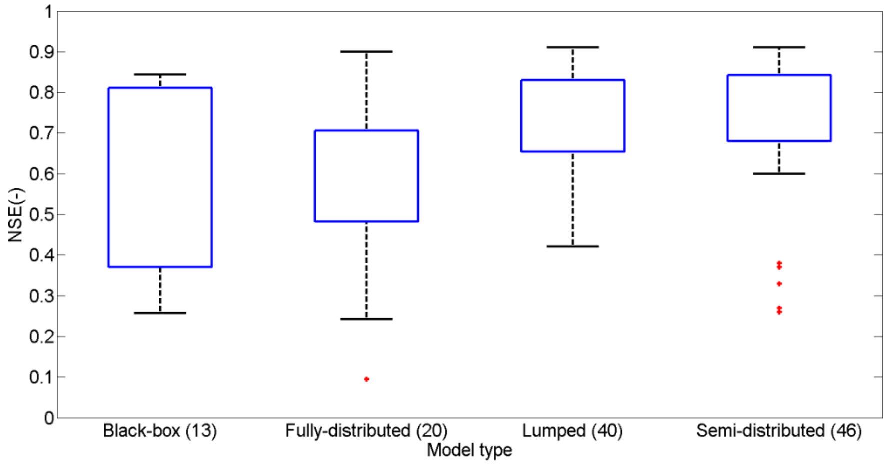

Figure 2. NSE of flow modelling stratified by model types (Level 1). The mean modelling results for each model type are 0.49, 0.58, 0.71 and 0.61 from the left to the right respectively. The number of catchments used for each model type is shown in parentheses on the x-axis tick labels. The boxes indicate 25-75% percentiles. The red dots present outliers.

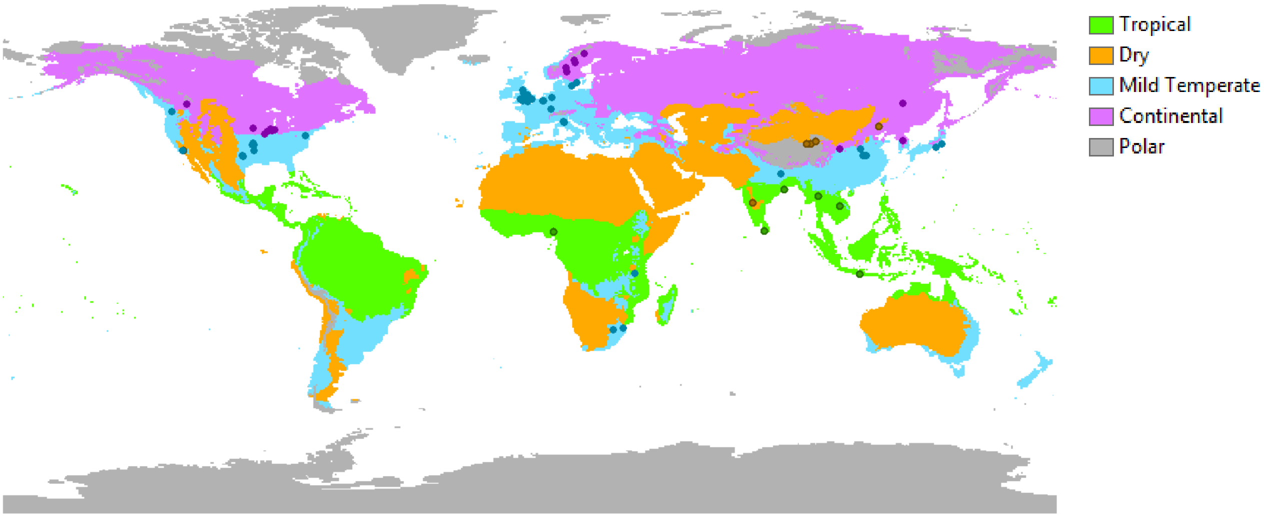

Figure 3. Climate map representing the catchment locations included in Level 1 meta-analysis.

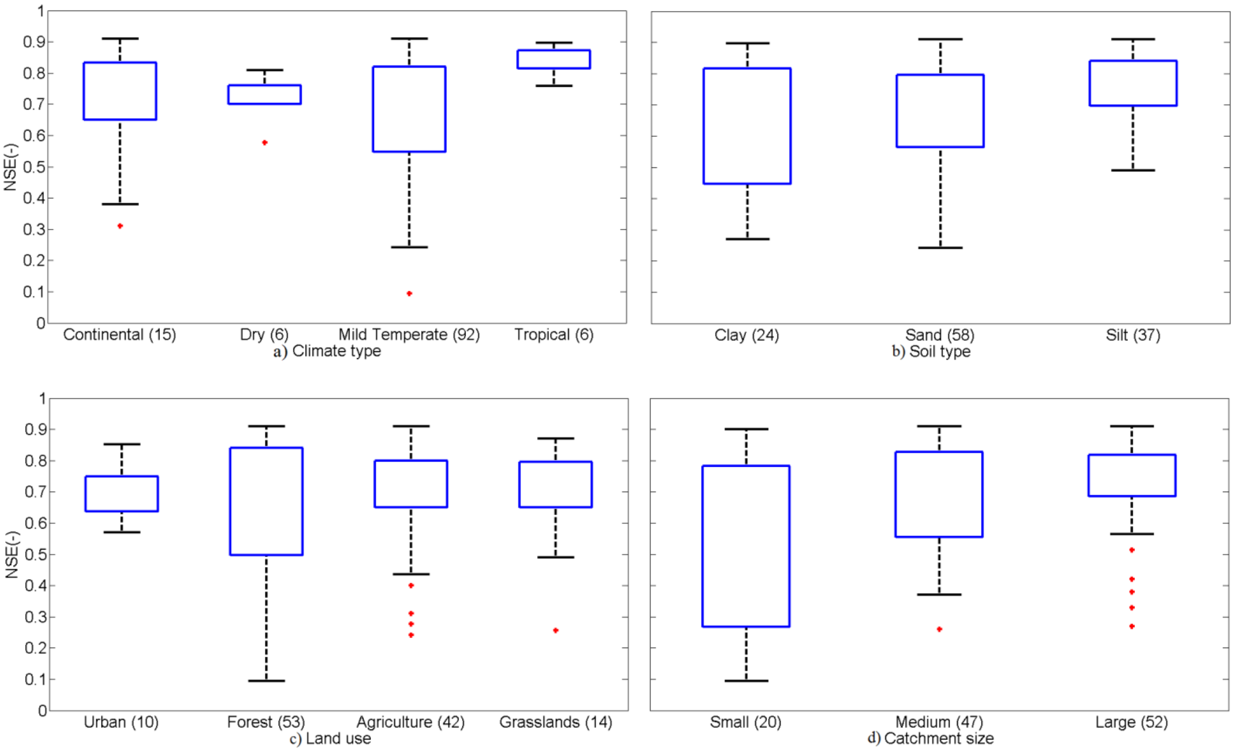

Figure 4. The NSE results of flow modelling, stratified by CSFs (Level 1). a) Climate regions: the mean simulation results for individual climate types are 0.71, 0.72, 0.59 and 0.84 from continental climate to tropical climate; b) Soil type: the mean simulation results for clay, sand and silt are 0.60, 0.53 and 0.76 respectively; c) Land cover: the mean simulation results for urban, forest, agriculture and grasslands are 0.70, 0.56, 0.66 and 0.70 respectively; d) Catchment size: The mean simulation results for small (0-100km^2), medium (100-1000km^2), and large (>1000km^2) catchments are 0.39, 0.63 and 0.71 respectively. The number of catchments used for each CSF group is shown in parentheses on the x-axis tick labels. The boxes indicate 25-75% percentiles. The red dots represent outliers.

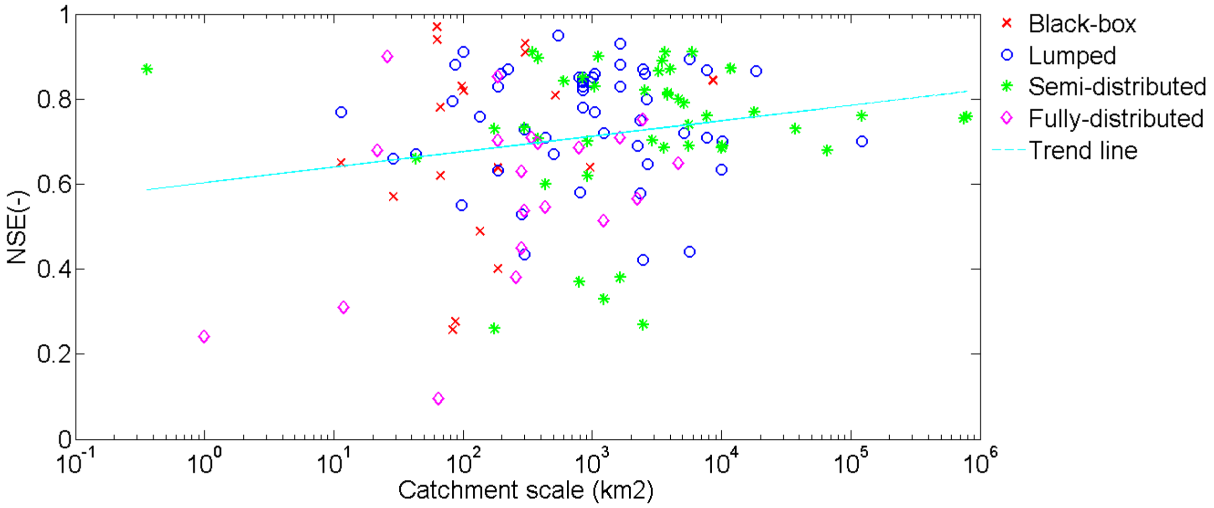

Figure 5. Performances across all model types with catchment scales in log. The trend line indicates the increase of model performance with the rise of catchment area.

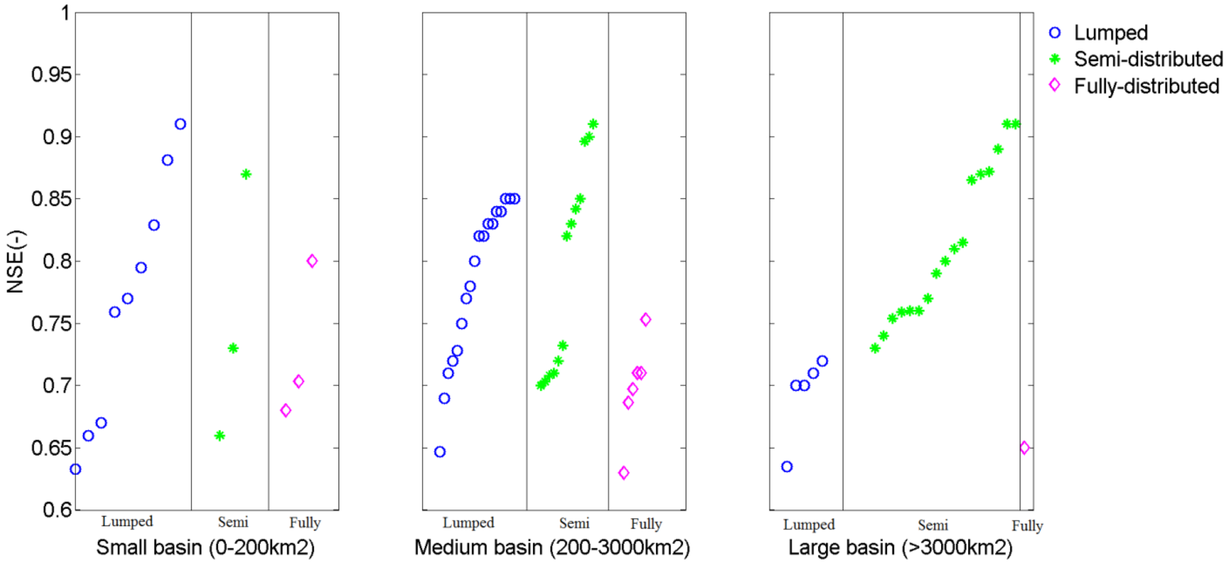

Figure 6. Relationship between model types and performances of flow modelling in respect to the most suitable border lines for small, medium and large catchments. It is important to note that within each catchment size group, the results are not shown in the order of catchment scales; instead they are sorted in the order of ascending NSE. The number of studies for black-box models is too small for comparison, so these results are excluded. Also it is noted that only the NSE results above 0.60 are shown here for a better visual illustration (Level 2).

Table 1. Summary assessment of hydrological models. Due to the limitation of paper layout, soil type and land cover are not shown here. In climate column, HC=Humid continental; HS= Humid subtropical; O=Oceanic; S=Subarctic; M=Mediterranean; SA=Semi-arid; TS=Tropical savanna; TR= Tropical rainforest; D= Desert. All studies with NSE performances less than 0 are not included in Figures. 2, 4, 5, 6 for better visual illustration.

Study	Region	Catchment area (km^2)	Climate	Model	Model type	Efficiency (NSE)
1. Chen et al., (2013)	China	121000	SA	Xinanjiang, HBV, TOPMODEL,	M2, M3	0.70, 0.76
2. Xie et al., (2013)	Midwest USA	17879-65742	HC	SWAT/HSPF	M3	0.65-0.73
3. He et al., (2012)	China	10000	HC	Xinanjiang, Sacramento, HBV, TOPMODEL	M2, M3	0.64, 0.69
4. Shi et al., (2011)	Mideast China	2550-10191	HS	Xinanjiang, SWAT	M2 , M3	0.70-0.86, 0.69-0.82
5. Pechlivanidis et al., (2010)	UK	10-1040	O	PDM	M2, M3	0.67-0.77, 0.66-0.83
6. Im et al., (2009)	Korea	258	S	MIKE SHE	M4	0.38
7. Zhang et al., (2008)	Northwest China	12	HC	MIKE SHE	M4	0.31
8. Cuo et al., (2008)	USA (Washington)	22	M	DHSVM	M4	0.68
9. Takeuchi et al., (2008)	China, Nepal, Sri Lanka, Japan, Southeast Asia,	603-795500	SA, TS, TR, HS	BTOPMC	M3	0.70-0.87
10. Cabus (2008)	Belgium	800	O	PDM	M2	0.58
11. Vischel et al., (2008)	South Africa	4625	O	TOPKAPI	M4	0.65
12. Rouhani et al., (2007)	Belgium	383	O	SWAT, MIKE SHE	M3, M4	0.71, 0.70
13. Singh et al., (2005)	USA (Illinois, Indiana)	5568	HC	SWAT, HSPF	M3	0.74, 0.69
14. Senthil et al., (2005)	Southeast India	515, 8570	SA, TS	ANN	M1	0.81, 0.84
15. Ackerman et al., (2005)	USA (California)	286, 338	M, SA	HSPF	M4	0.63, 0.71
16. Xiong et al., (2004)	China (Henan)	2623	HS	TOPMODEL	M2	0.80
17. Smith et al., (2004)	USA (Illinois)	65-2484	HS	SAC-SMA, SWAT, MIKE SHE, HRCDHM, HL-RMS	M2, M3, M4	-0.26-0.87, -2.58-0.60, 0.10-0.75
18. Saleh et al., (2004)	USA (Texas)	175, 921	HS	SWAT, HSPF	M3	0.26-0.62, 0.70-0.73
19. Whitaker et al.,(2003)	British Columbia	26	S	DHSVM	M4	0.90
20. Campling et al., (2002)	Nigeria	379	TS	TOPMODEL	M3	0.90
21. Bell et al., (2001)	UK	11-298	O	TCM, SAC-SMA, PDM, MCRM, IEM, TF, PRTF, Grid model,	M2, M1, M4	0.44-0.91, -0.85-0.82, 0.54-0.85

22. Tokar et al., (2000)	USA (Iowa, Maryland)	98, 960	HC, HS	ANN, SAC-SMA, SCRR	M1, M2	0.64-0.83, 0.84, 0.55
23. Yew et al., (1997)	Swaziland, Tanzania, USA (Oklahoma)	2344-2682	HS, O, D	Pitman, Sacramento, NAM, Xinanjiang, SMAR	M2	0.42-0.65
24. Xevi et al., (1997)	Germany	1	O	MIKE SHE	M4	0.24
25. Lindström et al., (1997)	Sweden	343-5975	S, O	HBV	M3	0.80-0.91
26. Todini (1996)	Italy	840, 4000	O	ARNO, TOPMODEL, Xinanjiang, Stanford IV, Sacramento, Tank, APIC, SSARR	M2, M3	0.78-0.87
27. Ambroise et al., (1996)	France	0.36	O	TOPMODEL	M3	0.87
28. Khan (1993)	USA (Oklahoma)	2344	D	Xinanjiang	M2	0.75

Table 2. Summary of climate types based on catchment locations and climate features, with corresponding major climate groups used for Level 1 meta-analysis, based on the Köppen climate classification system.

Climate types	**Climate main groups**
Humid continental, subarctic,	Continental
Desert, Semi-arid	Dry
Humid subtropical, oceanic, Mediterranean	Mild temperate
Tropical savanna, tropical rainforest,	Tropical

Table 3. The preferred and non-preferred (with 0.80 NSE as threshold) models used for small (0-200km^2), medium (200-3000km^2) and large (>3000km^2) catchments. The model names in the table are ordered from the most occurred to the least occurred, with the occurrence shown in brackets. When two models have the same number of occurrence, the one with lower NSE shows first in the non-preferred group, and the model with higher NSE is listed first in the preferred group (level 2). *PDM in small catchments is used as a semi-distributed model; TOPMODEL in medium catchments is used as a lumped model; ARNO in large catchments is used as a semi-distributed model

Models	**Small catchment**		**Medium catchment**		**Large catchment**
	M2	*M3*	*M2*	*M3*	*M3*
Preferred (NSE>0.8)	MCRM$_{(4)}$ SAC-SMA$_{(1)}$ IEM$_{(1)}$	TOPMODEL$_{(1)}$	SAC-SMA$_{(4)}$ Xinanjiang$_{(2)}$ Tank$_{(1)}$ Stanford$_{(1)}$ SSARR$_{(1)}$	HBV$_{(2)}$ TOPMODEL$_{(2)}$ SWAT$_{(1)}$ BTOPMC$_{(1)}$	HBV$_{(4)}$ BTOPMC$_{(3)}$ ARNO*$_{(1)}$
Not preferred (NSE<0.8)	SAC-SMA$_{(2)}$ PDM$_{(1)}$	PDM*$_{(1)}$ HSPV$_{(1)}$	SAC-SMA$_{(4)}$ PDM$_{(2)}$ Xinanjiang$_{(1)}$ APIC$_{(1)}$ TOPMODEL*$_{(1)}$	SWAT$_{(3)}$ HSPF$_{(1)}$ BTOPMC$_{(1)}$	SWAT$_{(5)}$ BTOPMC$_{(2)}$ HSPF$_{(1)}$ TOMODEL$_{(1)}$

Table 4. Catchment indicator Ψ for all the cases included in this study, in the format of {CSF, M} (i.e. four catchment significant features: catchment scale, climate region, soil type and land cover; and model type). The study no. is identical to the study no. in Table 1. It is noted that due to page limitation NSE results below 0 are not shown in this table. SI1, SI2, SI3 are small, medium and large catchments respectively; C1, C2, C3, C4 are continental, dry, mild temperate and tropical climates respectively; S1, S2, S3 are clay, sand and silt soil respectively; L1, L2, L3, L4 are the land covers for urban, forest, agriculture and grasslands respectively; M1, M2, M3, M4 are black-box, lumped, semi-distributed and fully-distributed models respectively.

Study no.	Case no.	Ψ code	Study no.	Case no.	Ψ code
1	1	SI3,C2,S2,L4,M3	18	41	SI2,C3,S2,L4,M3
1	2	SI3,C2,S2,L4,M2	19	42	SI1,C1,S2,L2,M4
2	3	SI3,C1,S3,L3,M3	20	43	SI2,C4,S1,L2,M3
2	4	SI3,C1,S3,L1,M3	21	44	SI1,C3,S2,L4,M2
3	5	SI3,C1,S3,L1,M2	21	45	SI1,C3,S2,L4,M1
3	6	SI3,C1,S3,L1,M3	21	46	SI1,C3,S2,L1,M2
4	7	SI3,C3,S3,L3,M3	21	47	SI1,C3,S2,L1,M1
4	8	SI3,C3,S3,L3,M2	21	48	SI1,C3,S2,L3,M2
4	9	SI2,C3,S1,L3,M3	21	49	SI1,C3,S2,L3,M1
4	10	SI2,C3,S1,L3,M2	21	50	SI1,C3,S3,L2,M2
5	11	SI1,C3,S2,L3,M2	21	51	SI1,C3,S3,L2,M1
5	12	SI1,C3,S2,L3,M3	21	52	SI1,C3,S3,L4,M2
5	13	SI2,C3,S2,L3,M2	21	53	SI1,C3,S3,L4,M1
5	14	SI2,C3,S2,L3,M3	21	54	SI1,C3,S1,L3,M2
6	15	SI2,C1,S1,L2,M4	21	55	SI1,C3,S1,L3,M1
7	16	SI1,C1,S2,L3,M4	21	56	SI1,C3,S1,L3,M4
8	17	SI1,C3,S2,L1,M4	21	57	SI1,C3,S1,L1,M2
9	18	SI3,C3,S3,L3,M3	21	58	SI1,C3,S1,L1,M1
9	19	SI3,C4,S3,L3,M3	21	59	SI1,C3,S1,L1,M4
9	20	SI3,C2,S2,L4,M3	21	60	SI2,C3,S1,L3,M2
9	21	SI2,C4,S1,L3,M3	21	61	SI2,C3,S1,L3,M1
9	22	SI2,C3,S2,L3,M3	21	62	SI2,C3,S1,L3,M4
9	23	SI3,C4,S1,L4,M3	22	63	SI2,C1,S3,L2,M1
10	24	SI2,C3,S2,L3,M2	22	64	SI2,C1,S3,L2,M2
11	25	SI3,C3,S1,L3,M4	22	65	SI1,C3,S2,L2,M1
12	26	SI2,C3,S2,L3,M3	22	66	SI1,C3,S2,L2,M2
12	27	SI2,C3,S2,L2,M4	23	67	SI2,C3,S2,L2,M2
13	28	SI3,C1,S3,L3,M3	23	68	SI2,C3,S1,L2,M2
14	29	SI2,C2,S1,L3,M3	23	69	SI2,C2,S3,L3,M2
14	30	SI3,C4,S3,L3,M3	24	70	SI1,C3,S2,L3,M4
15	31	SI2,C3,S3,L2,M4	25	71	SI3,C3,S2,L4,M3
15	32	SI2,C2,S3,L1,M4	25	72	SI3,C1,S2,L2,M3
16	33	SI2,C3,S3,L3,M2	25	73	SI2,C3,S2,L2,M3
17	34	SI1,C3,S1,L2,M2	25	74	SI3,C3,S2,L3,M3
17	35	SI1,C3,S1,L2,M3	26	75	SI3,C3,S3,L2,M3
17	36	SI1,C3,S1,L2,M4	26	76	SI2,C3,S3,L2,M3
17	37	SI2,C3,S1,L2,M2	26	77	SI2,C3,S3,L2,M2
17	38	SI2,C3,S1,L2,M3	27	78	SI1,C3,S3,L4,M3
17	39	SI2,C3,S1,L2,M4	28	79	SI2,C3,S3,L1,M2

18	40	SI1,C3,S2,L2,M3

Table 5. The preferred model types for each catchment situation are presented (i.e. model types with higher NSE accuracy). Where SI1, SI2, SI3 are small, medium and large catchments respectively; C1, C2, C3, C4 are continental, dry, mild temperate and tropical climates respectively; S1, S2, S3 are clay, sand and silt soil respectively; L1, L2, L3 are urban, forest, agriculture and grasslands lands respectively; M1, M2, M3, M4 are black-box, lumped, semi-distributed and fully-distributed models respectively.

Small catchment (CSFs)	**Model Type**	**Medium catchment (CSFs)**	**Model Type**	**Large catchment (CSFs)**	**Model Type**
SI1,C1,S2,L4	M4	SI2,C1,S3,L2	M2	SI3,C1,S2,L2	M3
SI1,C3,S1,L1	M4	SI2,C3,S1,L3	M2	SI3,C3,S2,L3	M3
SI1,C3,S2,L2	M1	SI2,C3,S2,L2	M3	SI3,C3,S2,L4	M3
SI1,C3,S2,L3	M2	SI2,C3,S3,L2	M2	SI3,C3,S3,L2	M3
SI1,C3,S3,L2	M2	SI2,C3,S3,L2	M3	SI3,C3,S3,L3	M3
SI1,C3,S3,L4	M3	SI2,C3,S3,L3	M2	SI3,C4,S1,L4	M3
		SI2,C4,S1,L3	M3		